

Testing a methodological framework for retrieving parallel texts in the domain of business translation

Daniel Gallego-Hernández

Departamento de Traducción e Interpretación. University of Alicante

Email: daniel.gallego@ua.es

Address: University of Alicante. Departamento de Traducción e Interpretación.
Carretera San Vicente del Raspeig s/n. 03690 San Vicente del Raspeig – Alicante
SPAIN

Daniel Gallego-Hernández is an associate professor in the Faculty of Arts of the University of Alicante in Spain. His research interests include corpus linguistics applied to translation, business translation and translation teaching. His PhD was entitled “Business translation and parallel texts on the internet”. He teaches business translation between French and Spanish as well as terminology at his faculty. His publications as an individual include various papers and a book entitled “Business translation and corpora”.

Testing a methodological framework for retrieving parallel texts in the domain of business translation

Daniel Gallego-Hernández

University of Alicante

Parallel texts are a kind of documentary resource that complements others, such as dictionaries, glossaries and terminology databases. Translators have used them for a long time and scholars have been discussing them since the last century. However, there seems to be a lack of consensus about both how to conceive of and how to use them. Against this backdrop, we propose a methodological framework that systematises the retrieval of parallel texts with search engines. First, we review some studies on the concept of parallel texts and others on the use of search engines in translation. Next, we put forward the actual framework. We then look at a series of case studies in which we apply our model to ten texts from the field of economics and business. The aim is to assess the results Google provides in response to a series of search queries. The results obtained show how useful a detailed knowledge of the texts to be retrieved and their possible location on the web is. Additionally, they can be considered for web for/as corpus methodologies.

Keywords: business translation; parallel texts; search engines, search queries

1. Definitions of parallel texts

Scholars have been discussing parallel texts and taking different approaches to them since the last century. Vinay and Darbelnet (1977, p. 272) distinguish between two types of parallel texts. The first deals with the same topic as the source text and involves identical or comparable situations. The second has a similar style to the source text and refers to a comparable situation.

Muñoz & Sánchez (1995, p. 170) combine the two types in question to conceive of parallel texts as texts written by different authors from different cultures but corresponding to similar communicative situations. On that basis, parallel texts deal with the same topic as the source text and present it in a similar style.

However, Lvovskaya (1992, p. 165) believes that a parallel text must not only deal with the same subject and area of knowledge as the source text, but also share its semantic and logical structure, content and textual function.

Nord (2005, p. 171) differentiates between parallel texts and model, comparative and background texts, which are also considered information resources:

“parallel texts” (i.e. TL texts on the same subject matter, belonging to the same genre as the ST), “model texts” (same subject matter, same genre, and even same genre variety, which allows the parallel text to be used as a model in lexis, sentence structures, and register features), “comparative texts” (same subject matter, different genre), or “background texts” (texts containing background information on the subject matter, e.g. encyclopaedic texts).

From a comparable viewpoint, Tricás (2003, p. 23) states that parallel texts are similar to the source text and written in the target language, the rules and conventions of which help translators identify. She also distinguishes between two other text types based on their potential for exploitation when translating, namely documentary texts, which help to better understand the subject and become familiar with its lexicon, and ready-made translations, which help to contrast originals with translations and to evaluate the gains and losses that arise when translating.

Mayoral (1996, p. 10-11), meanwhile, proposes a classification of resources based on two levels: (1) the textual level, which comprises multilingual texts, parallel texts written in the target language (authentic texts, forms and translated texts), and texts similar to the source text; and (2) the lexical level, which includes dictionaries, glossaries, specialised terminologies, legislation, manuals, brochures and informants. He states that parallel texts are target culture texts used for the same purpose as the source text, and recognises that the degree of parallelism between the source text and parallel texts is not always absolute (1996, p. 8).

Acuyo (2005, p. 253) says that there is not always a bijective correspondence between two texts from different cultures. Therefore, there may be cases where no parallel text exists in the target culture for certain communicative situations.

Sánchez (2003, p. 246-247) seems to conceive of parallel texts in the broad sense of the term and considers them from the angle of translators' needs. She affirms that their use should enable translators to obtain factual knowledge to help them understand the source text, and linguistic knowledge to help them produce the target text. She also states that the source text and a parallel text might not involve the same pragmatic and communicative situation.

Based on this overview of scholars' ideas, we can assume that there is consensus that parallel texts are useful for meeting information needs when translating (text-type conventions, terminology, phraseology, etc.). However, some scholars seem to conceive of parallel texts imprecisely. The label "similar texts", for example, is somewhat vague. Additionally, confusion has arisen in Translation Studies due to the existence of other terms imported from Corpus Linguistics, such as *parallel corpus*, i.e. texts in one language and their translations into another, and *comparable corpus*, i.e. original texts in two or more languages. In Translation Studies it is also possible to conceive of parallel texts as texts in one language and their translations into another. Some scholars, meanwhile, regard neither original texts and their respective translations nor texts that share certain characteristics as parallel texts. While some consider parallel texts to be solely target language texts dealing with the same situation, subject, etc., as a source text, and thus distinguish between them and other text types (forms, models, etc.), others (whom we agree with in this paper) are more flexible and recognise different degrees of parallelism between subjects or communicative situations.

2. Use of search engines

Search engines are one type of tool for looking for texts. Some scholars feel that translators use them differently from common users. Zanettin (2002, p. 241) states that some search engines “display, next to document pointers (hyperlinks), a concordance-like context with the search word(s) highlighted”. In this sense, the web can be considered a corpus, as in the “web as corpus” or “web as corpus surrogate” concepts (Bernardini, Baroni, and Evert, 2006, p. 10-11). The options search engines offer allow translators to use them for several purposes. For example, Gomez (1999) shows that simple searches enable translators to understand how terms are used in certain contexts. Simon and Swalef (2001, p. 245) say that the meaning of an abbreviation can be discovered by typing the word *abbreviation* followed by the abbreviation itself into Google’s search field. Grauwinkel and Neunzig (2004, p. 184) state that translators can query the web by typing in a term whose meaning they are looking for plus a keyword representing a lexicographical source to retrieve dictionary entries that define and explain the term.

Another “trick” identified by scholars is using search engines as if they were bilingual dictionaries. Campos (2007, p. 63) remarks that typing a term or specialised phraseological unit in one language along with a term likely to occur in the same context in another language (e.g. *declaración de obra nueva* in Spanish and *building* in English) into Google’s search field retrieves pages that contain both units, such as glossaries or dictionaries.

The web can also be used as if it were a concordancer, for the purpose, for instance, of confirming or rejecting hypotheses, checking frequencies or understanding the meaning of terms. Lindquist (1999, p. 188), according to whom “quite a lot of information can be gleaned from the lists of hits, without going to the actual sites”,

shows how to use simple searches to look for different English denominative variants of the term *osteoporosis*. Gomez (1999) states that translators can check the frequency of terms when they need to choose between two or more translation solutions, but also warns that caution must be exercised in relation to the number of results search engines obtain. Wallace (2001) refers to the possibility of using search engines as if they were parallel corpora or translation memories, and believes that multilingual websites are a useful resource as they provide translations of terms and longer segments.

However, some scholars are reluctant to use search engines as if they were concordancers. Perrotti (2005, p. 64) asserts that they do not always provide qualitatively adequate results and do not ensure a correct decision solely on the basis of numbers of results, and points out that the web contains all kinds of texts (translations, texts with spelling and/or grammatical errors, etc.). Bernardini et al. (2006, p. 10-11) consider that “the search is often time consuming, the relevance and authoritativeness of the solutions found is hard to assess, and the observation of recurrent patterns very difficult”. Kilgarriff and Grefenstette (2003, p. 345), meanwhile, state that search engines can frustrate translators when used as linguistic resources, since their results do not present enough instances or context, do not allow searches to be specified according to linguistic criteria, and give unreliable statistics.

We have seen that scholars refer to using search engines to consult parallel texts, but none of them suggests a framework that could help systematise this task. Some of them merely give examples of how to look for a translation equivalent or a text. Others seem reluctant to use the web as if it were a corpus, perhaps because of reservations about using search engines as if they were concordancers or the results that search engines provide. Against this backdrop, we believe that a methodological framework

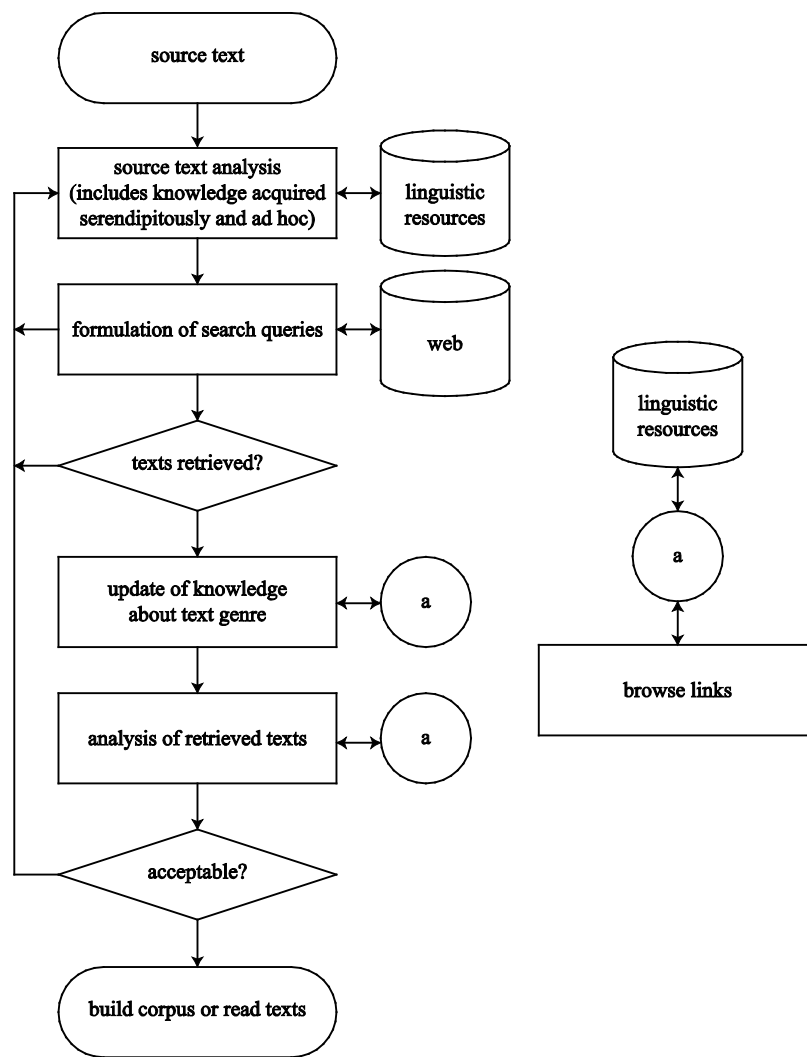
that can help both translators and translator trainers when retrieving parallel texts is required.

3. Presentation of the methodological framework

In this section, we propose a framework for using search engine queries to retrieve parallel texts. The model is based on two textual levels, hypertext and genre. In the former case, it essentially involves the structure and content of websites and webpages. In the latter case, it revolves around elements that can help identify genres, i.e. genre name, communicative situation, and structure. Search queries can be formulated on the basis of one or more aspects of both levels.

3.1. Sequence

Flowchart 1 shows a possible sequential process for retrieving texts from the web. It emphasizes both source and parallel text analysis with the help of linguistic resources, such as dictionaries.



Flowchart 1: The sequential framework

The process begins with source text analysis. Translators use their own knowledge (text analysis competence) and linguistic resources that can help them understand the text. This analysis, which can focus on aspects of hypertext and/or genre (see sections 3.2 and 3.3), should lead to the formulation of a query designed to retrieve parallel texts from the web via a search engine (see section 3.4).

If the search engine does not provide results, translators return to the previous stages to discover why. They can check whether they formulated the query correctly or reanalyse the text. If the search engine does provide results, translators update their knowledge about the source text while analysing the results to determine if they are acceptable. This analysis can focus on identifying the degree of parallelism between the

source text and the texts retrieved. To that end, translators can access the latter and consult relevant linguistic resources.

If the results are not acceptable (i.e. if the texts retrieved mostly deal with a subject or genre other than that of the source text, contain excessive noise, are of insufficient quantity, etc.), translators return to the previous stages of the process. If the results are acceptable, they can proceed to build their own *ad hoc* corpus (web for corpus) or read the texts retrieved (web as corpus) in order to meet their information needs when translating the source text.

3.2. Hypertext

Websites and webpages are two ways of structuring information on the web. Webpages are usually organised into directories or folders located on a server. A website is a set of webpages hierarchically linked to a homepage, identifiable by a URL and forming a recognisable documentary unit separate from other websites (Aguillo, 1998).

Websites can contain useful parallel texts for business translators. Alexander and Tate (2005) distinguish between different kinds of websites, among which business/marketing, informational and news websites are of particular note as far as relevance to economic and/or business information is concerned.

Business/marketing websites aim to sell a product or service. Most of them belong to commercial companies and have *com* domains (although some have geographic domains, such as *fr* or *es*). They can be divided into corporate sites, which provide shareholders, potential investors, the specialised press and/or society with information related to a company's situation, its corporate governance (articles of association, general meetings, etc.) and its finances (results, annual accounts, etc.), as well as general information on the company (presentation, history, etc.), press releases,

etc.; and commercial sites, which promote products and services to attract customers and therefore use commercial language.

News websites provide information on current affairs and are normally run by news organisations or other press-related bodies. In the field of economics, there are several (general or specialised) sites that can be useful sources of texts.

Informational websites usually present information on a particular topic. In the field of business and economics, there are various types of sites (administrative, institutional, legal, educational, financial, etc.) that contain a number of kinds of information, including reports on economic activities, regulations and laws, educational texts, and economic and financial databases.

The webpages or files that form websites of the kinds in question are usually HTML documents, but can also include others, such as PDF and DOC files, depending on the genre involved. In relation to the retrieval of parallel texts, it is worth noting that, as we will see in section 3.3, search engines make it possible to search for keywords that appear in the titles of such pages or for specific document formats or kinds of sites.

Every webpage has a unique Uniform Resource Locator (URL), which contains useful structural and semantic information relating not only to the webpage itself but also to the website it is part of. The URL *www.mediaset.es/inversores/es/Estatutos-Sociales-aprobados-JG_MDSFIL20150415_0008.pdf*, for example, includes the following components:

- The name of the website's owner (*Mediaset*, a mass media company).
- The geographic domain *es*, which specifies that the site's server is located in Spain. The site's information will therefore mainly be in Spanish.
- The folder *inversores* (*investors*), the information in which will be relevant to investors.

- The folder *es*, suggesting that there may be others containing the same information in other languages (e.g. *en*, a folder containing English content).
- The PDF file *Estatutos-Sociales-aprobados-JG_MDSFIL20150415_0008.pdf*, revealing that the URL provides direct access to *Mediaset*'s articles of association, as approved (possibly on 15.04.2015) at its annual general meeting (*JG* or *junta general*).

3.3. Genres

The notion of *genre*, the use of which is becoming increasingly common in translation and corpus-based studies, can also be a basis for formulating search queries. According to one of the most cited works on the subject (Swales, 1990, p. 58), genres belong to epistemological communities that share objectives and use a specific vocabulary acquired in educational or professional contexts. With regard to communicative purposes, a genre usually occurs in a conventional communication situation, targets a specific type of audience, and features similarities in terms of the structure, style and content of its exemplars.

Another definition, from the perspective of Translation Studies, is offered by García (2002, p. 3), who states that genres are conventionalised text forms that have specific functions in the culture they belong to and reflect senders' intended purposes, which receivers can anticipate. This suggests that a genre is a kind of container holding all the elements that help explain or identify it. Some of those elements can be employed when using search engines intelligently to retrieve texts.

Firstly, genre names can be used to formulate search queries. *Annual reports*, *annual accounts* and *bylaws* are examples of labels that can be used to retrieve online texts from the genres of the same names. To that end, the genre name must appear in a text so that the search engine can identify keywords and retrieve it. However, there are

genres the names of which do not appear in their texts, such as commercial correspondence or texts for selling products and services. Term variation poses an additional problem. For instance, if we wish to retrieve annual accounts, we should be aware that they may also be called *financial statements* or *financial reports*.

Closely connected to the names of genres are the main aspects of their communicative situations, such as field of expertise, mode, author and audience, and discourse. When the genre name is unclear or does not appear in parallel texts, it may be useful, as we will see in section 4, to consider the communicative situation involved and even relate its aspects to the types of websites mentioned earlier in this section.

Another key concept is *macrostructure*. According to Alcaraz (2000, p. 135), there are two types of macrostructure, one of which refers to the sections of genres' documents, and the other of which is related to components, known as "moves" (Swales, 1990), which make up each section, perform a specific function, do not necessarily match with paragraphs and can consist of lower units. For our purposes, web searches can be based on macrostructure to retrieve parallel texts containing specific blocks or segments of textual information. In practice, when a genre's macrostructure is clearly identified, texts can be retrieved by means of keywords included in the titles of their sections. When texts are not clearly structured, it is possible to work on the basis of moves and use certain elements of microtext as keywords. In that regard, parallel texts' microlinguistic lexical items, such as technical and semi-technical terminology, general vocabulary and specialised phraseology, can be taken into consideration.

In the case of business translation, it is helpful to know which genres translators work with. For our purposes, we use the taxonomy proposed by Herrero and Román (2015), which comprises the following five categories: (1) economic texts, which involve micro and macroeconomics (texts on public debt, press releases, dailies, etc.);

(2) financial texts, relating to financial operations and instruments, public debt and market analysis reports; (3) corporate texts, including texts with information on a company's activities and management (articles of incorporation, codes of conduct, management reports, etc.), documents containing its financial information (annual reports, stock option plans, credit ratings, etc.), contracts and agreements (confidentiality agreements, insurance contracts, etc.), texts about the company's operations, texts about business development, and texts that stem from consultancy and advisory services; (4) banking texts (retail and investment banking texts); and (5) accounting texts (audit reports, financial statements, etc.).

3.4. Search engines

To retrieve texts, it is not only necessary to analyse source and parallel text structure and their possible location on the web, but also to use search engines that provide quick access to them. One example is Google, the search engine most commonly used by translators.

Google's operators offer a range of options when looking for texts. For example, *site* can be used to search within a specific website or domain, or even in a particular folder within a site; *ext* retrieves documents with a specified file format; *intitle* searches for documents whose title contains a specified keyword; *inurl* locates files whose URL contains a specified keyword; and *related* looks for sites similar to a specified URL.

Google also supports other operators, such as the minus sign, which omits specified keywords from search results (ideal for disambiguation); quotation marks, which limit search results to texts that contain the exact string specified; and the asterisk, which acts as a placeholder for any word in an expression in quotation marks.

As can be inferred, these operators are closely linked to both hypertext and genre, making it possible to filter the search engine's results so as to retrieve a set of parallel texts that meet specific conditions.

4. Testing the framework

In this section, we firstly formulate various search queries designed to retrieve parallel texts for a series of source texts, and specify the rationale behind them. We then comment on the results obtained and discuss them in relation to alternative queries in order to comparatively evaluate our proposed queries. Such contextualisation (rationale and evaluation) is necessary to show how the suggestions made in section 3 can be regarded as a model.

4.1. Formulation of search queries

Table 1 shows the categories (column 1) to which the selected texts¹ (column 2) belong, as well as each text's search query (column 3) and the basis for each query's formulation (column 4), as explained in section 3. It also shows the approximate number of results Google provided in response to each query in May 2016 (column 5). This number is not that which Google displayed initially (which appears in figures 1-10), but that obtained when viewing the actual results (by clicking on the various "o"s in "Goooooogle"), which is normally reduced (in inconsistent ways).

CATEGORY	TEXT	SEARCH QUERY	No.	BASIS	RESULTS
economic	economic bulletins	"boletín mensual" "banco central europeo" ext:pdf site:www.bde.es OR site:www.ecb.europa.eu	1	author, mode, genre name	121
	analysis of macroeconomic indicators	"el pib * ha"	2	phraseology	272
financial	public debt	tesoro site:www.tesoro.es/sites/default/files/deuda/	3	author, website structure	40

¹ For space-related reasons, we have selected two examples from each of the categories proposed by Herrero and Román (2015) (see Table 1), most of which, according to Tolosa (2014), are text types frequently translated from English into Spanish and vice versa in Spain.

	investment funds	site:www.bbvaassetmanagement.com/am/tlwgpub/informes/ "perfil de riesgo del fondo y del inversor"	4	author, website structure, terminology	39
corporate	articles of association	"i denominación, objeto, duración y domicilio" ext:pdf	5	genre macrostructure, mode	97
	notices of annual general meetings	intitle:"convocatoria junta general" ext:pdf	6	genre name, mode	84
banking	current account contracts	"cuenta corriente" "condiciones generales" ext:pdf	7	terminology, macrostructure, mode	389
	takeover bids	oferta site:www.cnmv.es/opas/	8	author, website structure	86
accounting	external audit reports	"hemos auditado" ext:pdf	9	phraseology, mode	430
	financial statements	"cuentas anuales e informe de gestión" site:cnmv.es/AUDITA	10	genre name, website structure	285

Table 1: Search queries

Query 1 combines different aspects of hypertext and genre to retrieve economic bulletins. It specifically seeks documents from the websites of two official institutions, the European Central Bank and the Bank of Spain, which publish a range of documents including such bulletins, annual reports and statistics. The query also includes the genre name (*boletín mensual*) and the author of the documents to be retrieved (*banco central europeo*). Additionally, since these documents are published as PDF files, we use *ext:pdf* to prevent noise.

Query 2 is based on a phraseological unit ("*el pib * ha*") often used in the analysis of macroeconomic indicators. In this case, we use the term *PIB* (*GDP*) plus the determiner *el* (*the*), the auxiliary verb *ha* (*has*), and a non-specified series of characters between *pib* and *ha*. This strategy has the advantages of avoiding noise, such as pages that define the term *pib* (which Google shows if the query consists solely of the term *pib*), and of retrieving texts related to the GDP of different countries or regions (*el PIB alemán*, *el PIB venezolano*, *el PIB andaluz*, etc.).

Query 3 is designed to retrieve texts on public debt (in the financial category). In this case, since Spain has an official institution (*Tesoro Público*) that deals with public

debt, among other things, it is possible to query Google by using the *site* operator and specifying the exact directory or folder that contains texts of the kind in question (*sites/default/files/deuda*).

Texts on investment funds are the target of query 4. It is based on a term often included in such funds' prospectuses, namely *perfil de riesgo del fondo y del inversor* (*fund and investor risk profile*), and on the structure of the website of a bank that acts as an intermediary, www.bbvaassetmanagement.com/am/tlwgpúb/informes, within which *informes* (*reports*) is a folder containing texts on investment funds.

Query 5 retrieves articles of association and is based on their macrostructure, since they are clearly structured. The string selected to represent the macrostructure is *denominación, objeto, duración y domicilio* (*name, registered office, period of activity and objects*). Note that the letter *i* precedes this title, which can be called a *capítulo* (*chapter*) or a *título* (*title*). We omit these labels from the query to solve the problem of denominative variation they pose. Query 5 is also based on the format (PDF) in which articles of association are published on the web. Although they may sometimes be published in HTML format, we use *etx:pdf* to prevent possible noise.

Query 6 is *intitle:"convocatoria junta general" ext:pdf*. It is intended to retrieve notices of annual general meetings and is based on the name of the corresponding genre, *convocatoria junta general*. Since we expect this term to appear in the title of such documents, we use the operator *intitle*. Additionally, we use *ext:pdf* to look for PDF files.

Query 7 aims to retrieve account contracts, which are banking texts, on the basis of a subject-related term, *cuenta corriente* (*current account*), and macrostructure, as they often have clause titles that include the term *condiciones generales* (*general*

conditions). Since the query is not restricted to any particular bank, we expect to retrieve contracts from several financial institutions.

As suggested by Herrero and Román (2015), takeover bids can also be considered banking texts since banks sometimes act as intermediaries in such operations. As Spain's stock market regulator (*Comisión Nacional del Mercado de Valores*) has a repository that contains all kinds of documents, including takeover bids, query 8 uses the term *oferta* and the *site* operator to retrieve bid prospectuses.

Query 9 aims to retrieve accounting texts, specifically audit reports. We use a simple expression, *hemos auditado* (*we have audited*), which normally appears at the beginning of such documents. To avoid noise we also include *ext:pdf* in the query, since original audit reports are published as PDF files.

Query 10 is designed to retrieve annual accounts. As mentioned earlier, Spain's stock market regulator offers different kinds of documents related to companies involved in stock markets. As the structure of its website includes a folder containing financial statements, it is possible to use the *site* operator in combination with the genre name, *cuentas anuales* (*annual accounts*), which occurs together with the term *informe de gestión* (*management report*).

4.2. Discussion of results

Our case studies are designed for the purpose of assessing the results obtained by Google in response to the aforementioned queries. To that end, we take into account both the number of results displayed (see Table 1) and whether or not the texts retrieved are actually documents of the kind sought by each query (see annexes). We believe that analysing the first ten results displayed for each query can provide an approximate idea of the overall accuracy of all its results.

Nine of the first ten results obtained in response to query 1 link to European Central Bank economic bulletins published in different years. The results are from the websites of both the European Central Bank itself and the Bank of Spain. Only the eighth result is not an actual economic bulletin. The corresponding document comprises pages from section VIII of the bulletin published in December 2008, and appears among the results because it seems that the bulletins are also published in sections in separate files. The number of results obtained is quite large (121), especially considering that the search was restricted to just two sites and that the quantity of bulletins is limited. It is not the case that we retrieved 121 distinct bulletins. Some files could be duplicates, since we were searching two sites containing the same documents, and, as stated, some results link to sections rather than full bulletins.

Query 2 returns 272 results, mainly because no filters were applied and the asterisk included in the phraseological unit allows for numerous alternatives. Six results are from the websites of Spanish newspapers (*elmundo.es*, *20minutos.es*, *diariovasco.com*, *libremercado.com*, *bolsamania.com* and *economia.elpais.com*). Three others link to personal webpages, specifically *malaprensa.com*, which is a blog on mistakes found in the press, *dlacalle.com*, which belongs to an economist who also publishes in the press, and *economiaparapincipiantes.com*, which belongs to economics, business and management graduates. The first result corresponds to a specific department of the CaixaBank banking institution (*caixabankresearch.com*) the mission of which is to create and spread knowledge through economic research and analysis. Of the ten documents, only the one from *malaprensa.com* is not on macroeconomic indicators. The document from *dlacalle.com* contains some subjective elements since it involves the author's opinion on the Spanish GDP. The document from *economiaparapincipiantes.com* explains how to interpret news on GDP.

Query 3 provides 40 results due to being restricted to a specific server, *tesoro.es*. Five of the Spanish documents obtained (results 2, 3, 7, 9 and 10) are monthly bulletins about the central government debt market (market overviews, auctions, turnover, debt portfolio, etc.). The others provide statistics on various government fixed income securities.

The number of results (39) obtained in response to query 4 is also low, since the search was restricted to a particular financial institution's website (*bbvaassetmanagement.com*). The first ten results displayed all link to investment fund prospectuses.

Query 5 provides 97 results. Since it was based on the macrostructure of articles of association and required Google to retrieve texts containing the exact sequence *denominación, objeto, duración y domicilio* (name, registered office, period of activity and objects), the search engine did not retrieve texts containing variations of the specified sequence, such as *denominación, objeto, duración, domicilio*, where the word *y* (and) is missing; *denominación, objeto, duración, comienzo de las operaciones sociales y domicilio* or *denominación, objeto, duración, ámbito y domicilio*, where *comienzo de las operaciones* (start of operations) or *ámbito* (geographical location) is included between *duración* and *domicilio*; or *denominación, domicilio, objeto, duración y ejercicio*, where *objeto, duración* and *domicilio* appear in different positions. The first ten results link to articles of association of companies with different legal forms (e.g. public and private limited companies). One of them (result 6) links to a model, which is from an informational website (*uned.es*) rather than a commercial website.

Similarly, query 6, which obtains 84 results, does not retrieve documents with variations of the genre name, such as *convocatoria a junta general, convocatoria de junta general* or *convocatoria a la junta general*, where a preposition (*a* or *de*) is

included between *convocatoria* (notice) and *junta general* (annual general meeting). The first ten results link to notices of general meetings. Results 2 and 10 are not from commercial websites but informational websites related to finances, namely *bolsasymercados.es* and *bolsamadrid.es*. This suggests that these websites include repositories linking to corporate texts.

Query 7 was based on a subject-related term (*cuenta corriente*) and a term related to the macrostructure of all kinds of contracts (*condiciones generales*). This may partly explain why it provides 389 results. Another factor could be the search not being restricted to a particular site or domain. It is therefore possible, for example, to obtain results from different countries, including Uruguay (result 7) and Peru (result 10). As for the accuracy of the first ten results, not all the corresponding documents are specifically related to current accounts; some of them also deal with other kinds of accounts or means of payment. In any case, they are all contracts and can be used as parallel texts.

Query 8 yields a low number of results (86), seemingly due to it being restricted to a specific folder of the website *cnmv.es*. The first ten documents retrieved are all bid prospectuses and their filenames are *opa*.pdf*.

As in the case of query 2, the ninth is based on a phraseological unit, this time comprising two words (*hemos auditado*), and is not restricted to any given site. Consequently, Google obtains 430 results. The first results displayed link to audit reports. As is to be expected, bearing in mind the strategy used, they are not all from the same country (result 3 is from Mexico) or related to the same type of organisation. Most of them actually link to institutional organisations, foundations or associations rather than trading companies (a different strategy would be required to retrieve only trading companies' audit reports).

Google provides 285 results in response to query 10, which, like query 8, is restricted to a particular folder of the website *cnmv.es*, specifically that which contains financial statements, although on this occasion we also included a term representing the genre name, *cuentas anuales e informe de gestión* (*annual accounts and management report*). We assume that the folder contains other documents besides those retrieved by Google. Since the query is limited to the site of the Spanish stock market regulator, the first ten results obtained not only link to annual accounts but also to documents belonging to public limited companies, unlike in the case of our ninth query, which involved no restrictions.

So far, we have commented on the results obtained in response to each query. According to the framework presented in section 3, various strategies can be used, depending on which source text elements are considered when formulating queries. Table 2 contains alternative queries for comparatively evaluating those proposed originally.

SEARCH QUERY	No.	BASIS	RESULTS
“evolución económica y monetaria” ext:pdf	1a	macrostructure, mode	20
pib site:http://cincodias.com	2a	terminology, author	606
“bonos” “deuda del estado” “entorno macroeconómico” ext:pdf	3a	terminology, mode	40
“Nº Registro Fondo CNMV” OR intitle:”folleto simplificado” ext:pdf	4a	macrostructure, genre name, mode	76
“La duración de la Sociedad es” ext:pdf	5a	phraseology, mode	244
“orden del día” site:www.bolsamadrid.es/docs/hechos/	6a	macrostructure, author	249
“contrato regula” “titular de la cuenta” ext:pdf	7a	terminology, phraseology, mode	50
“Elementos subjetivos de la Oferta”/”Elementos objetivos de la Oferta” OR intitle:”oferta pública de adquisición “ ext:pdf	8a	macrostructure, genre name, mode	83
intitle:”informe de auditoría” ext:pdf	9a	genre name, mode	192
“se valora inicialmente por su” ext:pdf	10a	phraseology, mode	258

Table 2: Alternative search queries

Query 1a is based on the genre macrostructure. It retrieves some European Central Bank economic bulletins plus some parallel texts from the websites of both the European Central Bank itself and the Bank of Spain. It provides fewer results than query 1.

Query 2a is based on author. The website *cincodias.com* corresponds to a business journal and is a likely source of parallel texts. Query 2a gives more results than query 2, and also retrieves texts that analyse macroeconomic indicators.

Query 3a is based on the terms *bonos* (*bonds*), *deuda del estado* (*government debt*) and *entorno macroeconómico* (*macroeconomic environment*). It retrieves texts on public debt from several government servers, such as *tesoro.es*, *mineco.gob.es* (Spanish Ministry of Finance) and *cnmv.es*. It provides the same number of results (40) as query 3.

The alternative strategy query 4a is based on macrostructure, in the form of *Nº Registro Fondo CNMV* (*fund registration number in CNMV*), and on the genre name *folleto simplificado* (*simplified prospectus*). It yields 76 results, slightly more than query 4. All the texts retrieved are on investment funds and come from different websites.

Query 5a employs a genre-based strategy too. It uses the phraseological unit *la duración de la sociedad es* (*the company's period of activity is*). Like query 5, its first results link to different articles of association and models. It retrieves a higher number of texts (244) than query 5 (97).

Combining genre- and URL-based strategies, query 6a uses the term *orden del día* (*matters on the agenda*), as part of the macrostructure of notices of annual general meetings, and restricts the search to *bolsamadrid.es* (the Madrid Stock Exchange website), more specifically to the folder *docs/hechos*, which contains publications on *hechos relevantes* (*relevant facts*), i.e. any event or information that may have a discernible effect on companies' share performance. The query provides 249 results

(many more than query 6) related to annual general meetings (not only notices but also resolutions).

In response to query 7a, which is based on the phraseological unit *contrato regula* (contract governs) and the term *titular de la cuenta* (account holder), Google returns 50 results (far fewer than query 7). All the PDF documents retrieved link to models and contracts for different banking products, including current accounts.

In the case of query 8a, texts on takeover bids are retrieved using various genre-based strategies. The terms *elementos subjetivos de la oferta* and *elementos objetivos de la oferta* (objective and formal elements of the offer) are parts of the genre's macrostructure, while *oferta pública de adquisición* (takeover bid) is the main genre name. The number of results (83) is practically the same as query 8 provided. The texts retrieved are from different companies' websites and *cnmv.es*, and link directly to offer documents or changes to the terms of bids.

Query 9a is based on genre name. It provides 192 results (far fewer than query 9), most of which link to external audit reports or models.

The last query, 10a, is based on the phraseological unit *se valora inicialmente por su* (are initially recognised at). It returns 258 results (slightly fewer than query 10), which all link directly to annual accounts of different kinds of organisations (societies, groups, foundations, government institutions, etc.).

As a general appraisal, there is no real difference between the number of results obtained in response to each alternative query and to the corresponding query in Table 1. Additionally, looking at the different texts retrieved shows a similar degree of parallelism in both cases.

It can thus be said that both the original and the alternative queries worked well. While there are well known arguments against using Google counts in research, the

actual number of results displayed on the final results page and accessible to translators (an average of 180 per search query) seems appropriate for building a corpus or consulting texts. Also, it is possible to add to the results obtained initially by using different strategies to retrieve further parallel texts. In terms of the degree of parallelism of the texts retrieved, most of each query's first ten results were genuine parallel texts, showing that search engines can provide qualitatively adequate results when used strategically.

Additionally, the framework tested sheds light on the kinds of strategies that can be used when retrieving texts. They are mainly URL- and genre-based strategies, which can be combined according to the genre of the texts that translators need to retrieve and their knowledge about it. Genre-based strategies are useful when parallel texts are available from different servers or when translators do not know of any server potentially containing parallel texts. Such strategies can also be used to acquire and update specialised knowledge about a source text (e.g. kinds of servers liable to host parallel texts, the source text's specific and idiosyncratic terminology or phraseology, etc.). URL-based strategies are useful when parallel texts are normally published in a particular format or when translators know of a specific server containing them.

5. Conclusions

In this paper we have proposed and tested a framework for the retrieval of parallel texts in the domain of business and finance. The model presented involves analysing texts from two different perspectives (genre and hypertext) with a view to using search engines intelligently. Firstly, translators can consider a series of aspects related to text genres, such as their name, macrostructure, terminology and phraseology. Additionally, they should take into account how webpages and websites are structured (URL, site types, text titles).

Our case study results lead us to think that it is possible for translators to retrieve parallel texts closely related to a source text with a very high degree of accuracy and without noise. Once a search engine displays its results, translators, scholars and translator trainers alike can use them to view the corresponding texts (web as corpus) or even to download texts to build a corpus (web for corpus).

There are other strategies besides those we have used, and translators can consider possible terminological and/or structural variations involving microtext and/or macrostructure. Additionally, strategies can be improved, and translators should not expect to choose an optimal strategy from the outset. They can broaden their knowledge of the field involved on the basis of the first results they obtain (they might discover new text repositories or identify term variations, for example) and subsequently formulate new queries.

The results of this study are not definitive and our proposed framework should be tested with other languages or texts.

Bibliography

- Acuyo, M. (2005). El concepto de texto paralelo: algunas consideraciones para la traducción especializada. *Polissema*, 5, 247–262. Retrieved from http://www.iscap.ipp.pt/~www_poli/
- Aguillo, I. (1998). Hacia un concepto documental de sede web. *El Profesional de la Información*, 7, 1–2. Retrieved from <http://www.elprofesionaldelainformacion.com>
- Alcaraz, E. (2000). *El inglés profesional y académico*. Madrid: Alianza.
- Alexander, J., & Tate, M. (2005). Original Web Evaluation Materials. Retrieved from http://www.widener.edu/about/campus_resources/wolfgram_library/evaluate/original.aspx
- Bernardini, S., Baroni, M., & Evert, S. (2006). A WaCky Introduction. In M. Baroni, & S. Bernardini (Eds.), *WaCky! Working Papers on the Web as Corpus* (pp. 9–40). Bologna: Gedit.

- Campos, M. (2007). Documentación y traducción en el sector inmobiliario. In P. Fuertes (Ed.), *Problemas lingüísticos en la traducción especializada* (pp. 49–66). Valladolid: Universidad.
- García, I. (2002). El género: plataforma de confluencia de nociones fundamentales en didáctica de la traducción. *Discursos. Série Estudos de Tradução*, 2. Retrieved from <https://repositorioaberto.uab.pt/>
- Gómez, R. (1999). Internet en la traducción: ideas sobre el uso de Internet en la traducción del inglés al español. *Puntoycoma*, 57. Retrieved from <http://ec.europa.eu/translation/bulletins/puntoycoma/57/pyc577.htm>
- Grauwinkel, G., & Neunzig, W. (2004). El uso inteligente de las tecnologías de la información: un reto para la traducción especializada inversa. In R. Gaser, C. Guirado, & J. Rey (Eds.), *Insights into Scientific and Technical Translation* (pp. 175–185). Barcelona: Universitat Pompeu Fabra.
- Herrero, L., & Román, V. (2015). English to Spanish translation of the economics and finance genres. *Intralinea. Special Issue: New Insights into Specialised Translation*. Retrieved from http://www.intralinea.org/specials/article/english_to_spanish_translation_of_the_economics_and_finance_genres
- Kilgariff, A., & Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3), 333–348. doi: 10.1162/089120103322711569
- Lindquist, H. (1999). Electronic Corpora as Tools for Translation. In G. Anderman, & M. Rogers (Eds.), *Word, Text, Translation; Liber Amicorum for Peter Newmark* (pp. 179–189). Clevedon: Multilingual Matters.
- Lvovskaia, Z. (1992). Concepto de texto paralelo en la traducción especializada. In M. Edo (Ed.), *Actes del I Congrés Internacional sobre Traducció* (pp. 163–168). Barcelona: Universitat Autònoma.
- Mayoral, R. (1996). La traducción comercial: estrategias de traducción en razón a la disponibilidad de las fuentes de referencia terminológicas y textuales. *V Curso Superior de Traducción Inglés/Español*. Unpublished manuscript.
- Muñoz, R., & Sánchez, M. (1995). Textos paralelos. Proyecto de investigación. In R. Martín (Ed.), *V Encuentros complutenses en torno a la traducción* (pp. 169–174). Madrid: Universidad Complutense.
- Nord, C. (2005). *Text analysis in translation. Theory, methodology, and didactic application of a model for translation-oriented text analysis* (C. Nord, and P. Sparrow, Trans.) (2nd ed.). Amsterdam/Atlanta: Rodopi.

- Perrotti, A. (2005). Uso de Corpus Customizado como Fonte de Pesquisa para Tradutores. *Confluências*, 3, 62–79. Retrieved from http://www.scientiavinces.com/ana/confluencias%20n3_perrotti-garcia.pdf
- Sánchez, P. (2003). *Els documents digitals especialitzats: utilització de la lingüística de corpus com a font de recursos per a la traducció especialitzada* (Unpublished doctoral dissertation). Universitat Autònoma de Barcelona. Retrieved from <http://www.tesisenxarxa.net/handle/10803/5261;jsessionid=490E566DA0F190F7AB868F61189F5C85.tdx1>
- Simon, B., & Swalef, H. (2001). Stratégies de recherche sur l'Internet. *T&T*, 3, 244–250. Retrieved from <http://bookshop.europa.eu>
- Swales, J. (1990). *Genre Analysis*. Cambridge: University Press.
- Tolosa, M. (2014). Dime qué traduces y 'les' diré quién eres. Estudio basado en encuestas acerca de los documentos traducidos por traductores económicos (inglés-español y español-inglés). In D. Gallego (Ed.). *Traducción económica: entre profesión, formación y recursos documentales* (pp. 23-41). Soria: Diputación Provincial.
- Tricás, M. (2003). *Manual de traducción (francés-castellano)* (2nd ed.). Barcelona: Gedisa.
- Vinay, J., & Darbelnet, J. (1977). *Stylistique comparée du français et de l'anglais. Méthode de traduction* (2nd ed.). Paris: Didier.
- Wallace, J. (2001). Bilingual and multilingual websites as a translation resource. *The MICATA Monitor*, 25(2), 3–5.
- Zanettin, F. (2002). DIY Corpora: the WWW and the Translator. In B. Maia, J. Haller, and M. Ulrych (Eds.). *Training the Language Services Provider for the New Millennium* (pp. 239–248). Porto: Universidade.

Annexes

"boletín mensual" "banco central europeo" extpdf site:www.bde.es OR site:www

Todo Noticias Imágenes Vídeos Maps Más ▾ Herramientas de búsqueda

Aproximadamente 1.200 resultados (0,27 segundos)

BOLETÍN MENSUAL. 10.º Aniversario BCE - European Central Bank
https://www.ecb.europa.eu/pub/pdf/other/10thanniversaryoftheecmb2_00806es.pdf?..
Boletín Mensual · 10.º aniversario del BCE. PRSLOGO. El Banco Central Europeo (BCE) se creó el 1 de junio de 1998, convirtiéndose en el banco central.

Boletín Mensual. Marzo 2014 - Banco de España
www.bde.es/f/webbde/SES/Secciones/.../BoletínMensual/BCE/14/Fich/bm1403.pdf
5 mar. 2014 · Publicaciones del Banco Central Europeo · VII ... La presente edición del Boletín Mensual contiene un artículo titulado «Proyecciones ...»

Boletín Mensual. Diciembre 2014 - Banco de España
www.bde.es/f/webbde/SES/Secciones/.../BoletínMensual/BCE/14/Fich/bm1412.pdf
3 dic. 2014 · El Boletín Mensual del BCE de diciembre de 2014 será el último, ya que será sustituido por un ... Publicaciones del Banco Central Europeo · VII.

Boletín Mensual. Marzo 2014. Publicaciones del Banco Central Euro
www.bde.es/f/webbde/SES/Secciones/.../BoletínMensual/BCE/14/Fich/bm1403dp.pdf
PUBLICACIONES DEL BANCO CENTRAL EUROPEO · VII. Boletín Mensual · BCE. Marzo 2014. El BCE publica distintos documentos que proporcionan ...

Boletín Económico. Junio 2015 - Banco de España
www.bde.es/f/webbde/SES/Secciones/Publicaciones/.../15Jun/Fich/be1506.pdf
19 jun. 2015 · FUENTES: Banco Central Europeo y Banco de España. a Fecha de cierre de elaboración de los supuestos: 15 de junio de 2015. 2015.

Boletín Mensual - Banco de España
www.bde.es/f/webbde/SES/Secciones/.../BoletínMensual/BCE/14/Fich/bm1406.pdf
4 jun. 2014 · Publicaciones del Banco Central Europeo · VII ... La presente edición del Boletín Mensual contiene un artículo titulado «Proyecciones ...»

Boletín Mensual. Diciembre 2013 - Banco de España
www.bde.es/f/webbde/SES/Secciones/Publicaciones/.../BoletínMensual/BCE/.../bm1312.p...
5 dic. 2013 · Publicaciones del Banco Central Europeo · IX. Glosario ... La presente edición del Boletín Mensual contiene un artículo titulado «Proyecciones ...»

Diciembre 2008. Documentos publicados por el Banco Central Europeo
www.bde.es/f/webbde/Secciones/Publicaciones/.../BoletínMensual/BCE/.../bm0812dp.pdf
Boletín Mensual · Diciembre 2008. VII. DOCUMENTOS PUBLICADOS POR EL BANCO CENTRAL EUROPEO · A PARTIR DEL 2007. Esta lista facilita ...

Boletín Mensual del BCE. Abril 2006 - European Central Bank
https://www.ecb.europa.eu/pub/pdf/other/pq75-86_mb200604es.pdf?..
5 abr. 2006 · Banco Central Europeo · 2006. Dirección. Apartado de ... La publicación impresa definitiva del Boletín Mensual del BCE en español puede ...

Informe anual 2010 - European Central Bank - Europa.eu
<https://www.ecb.europa.eu/pub/pdf/an/nrep/ar2010es.pdf?..>
Banco Central Europeo · 2011. Dirección ... 2010, en el artículo titulado «La respuesta del BCE a la crisis financiera», Boletín Mensual · BCE, octubre de 2010.

Figure 1

"el pib * ha"

Todo Noticias Imágenes Vídeos Maps Más ▾ Herramientas de búsqueda

Aproximadamente 22.900.000 resultados (0,65 segundos)

Sobre la medición y el uso del PIB | CaixaBank Research
www.caixabankresearch.com/y/4111m-d1-es
En la eurozona, el peso de la inversión sobre el PIB se ha reducido del 22% en 2007 hasta el 18% actual. En EE. UU. el panorama es similar. Ante estos datos ...

El turismo consolida su repunte por el tirón de la demanda interna ...
economia.elpais.com · Economía
12 ene. 2016 · El PIB turístico ha crecido un 3,7% respecto a 2014 tras recibir 68 millones de turistas este año. La gran mayoría de las empresas del ...

En las noticias

El PIB de Alemania crece un 0,7% durante el primer trimestre
elsalamanla.com · hace 1 día
El PIB alemán ha subido un 0,7% durante el primer trimestre con respecto al mismo periodo ...

Más noticias sobre "el pib * ha"

Razones para creer el PIB | dlacalle.com
www.dlacalle.com/razones-para-creer-el-pib/
20 jun. 2014 · ... desde 2007 la correlación previa entre indicadores industriales se ha roto y el PIB se ha moderado mucho menos que dichos indicadores.

El PIB de Euskadi creció en 2015 el 2,8%, el doble que en 2014 ...
www.diariorvasco.com/_/24/euskadi-cierre-2015-crecimiento-20160224110615.html
24 feb. 2016 · El PIB de Euskadi ha registrado un crecimiento del 3,1% en el cuarto trimestre de 2015 con respecto al mismo periodo del año anterior, según ...

¿Cómo interpretar una noticia del PIB? - Economía Para Principiantes
www.economia-paraprinipiantes.com/2014/12/como-interpretar-una-noticia-del-pib/
30 dic. 2014 · En este, el primer dato que nos dan (el 0,5%) se refiere a la tasa trimestral/trintertrimestral, eso quiere decir que el PIB trimestral ha crecido un 0 ...

Se mantiene el crecimiento en el PIB gracias a la ralentización en las ...
www.libremercado.com/_/se-mantiene-el-crecimiento-en-el-pib-gracias-a-la-ralentiza...
27 feb. 2016 · Destacar por otra parte que, en el conjunto del año, el PIB nominal ha crecido un 3,8% tal como estimaba el Gobierno. El problema surge ...

El PIB de Extremadura crece un 3 por ciento en 2015, ligeramente por ...
www.20minutos.es/_/pib-extremadura-crece-3-por-ciento-2015-ligeramente-por-deb...
30 mar. 2016 · El PIB de Extremadura ha crecido un 3 por ciento en 2015, ligeramente por debajo del 3,2 por ciento de incremento registrado a nivel nacional ...

Malaprensa: ¿El PIB o el crecimiento del PIB?
www.malaprensa.com/2015/10/el-pib-o-el-crecimiento-del-pib.html
19 oct. 2015 · Parece decir que el PIB se ha reducido hasta quedar en un 6,9 % del anterior. O sea una reducción del 93,1 %. El 6,9 % sería, no la reducción ...

La lucha anticorrupción hunde el PIB de Macao un 26,5% | Economía ...
www.elmundo.es · Economía
18 sept. 2015 · Desde la promoción del 'boom' de los casinos del año 2002 hasta el inicio de la campaña anticorrupción, el PIB de Macao ha venido ...

Figure 2



Figure 7

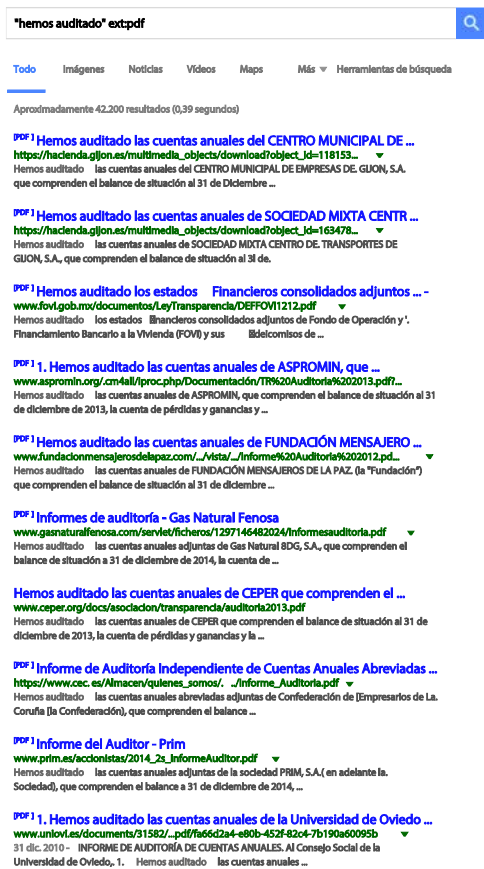


Figure 9



Figure 8

